Winning Space Race with Data Science

Youssef Shanan

3/6/2022

Outline

- Executive Summary
- Introduction
- Methodology
- Results
- Conclusion
- Appendix

Executive Summary

- Summary of methodologies
  - Data collection using SpaceX API and webscraping.
  - Data wrangling.
  - Exploratory data analysis using SQL and visualization.
  - Interactive visual analytics using Folium.
  - Interactive visual analytics and a dashboard using Plotly Dash.
  - Predictive analysis (Classification).
- Summary of all results
  - Using SQL, we discovered that the total number of successful and failed mission outcomes are 101.
  - Using Folium, we discovered that the nearest highway to CCAFS SLC-40 is 0.60 KM.
  - Using Plotly Dash, we discovered that KSC LC-39A had the most success launches.
  - Using predictive analysis, we discovered that the decision tree classifier resulted in the highest accuracy.

Introduction

- Project background and context
  - The aim of this project is to predict if the Falcon 9 first stage will land successfully. SpaceX advertises Falcon 9 rocket launches on its website with a cost of 62 million dollars because SpaceX can reuse the first stage. Therefore, if we can determine if the first stage will land, we can determine the cost of a launch. This information can be used because SpaceY wants to bid against SpaceX for a rocket launch.
- Problems you want to find answers
  - What are the main characteristics of a successful or failed landing?
  - What is the total number of successful launches according to launch site?

Section 1

Methodology

Methodology

Executive Summary

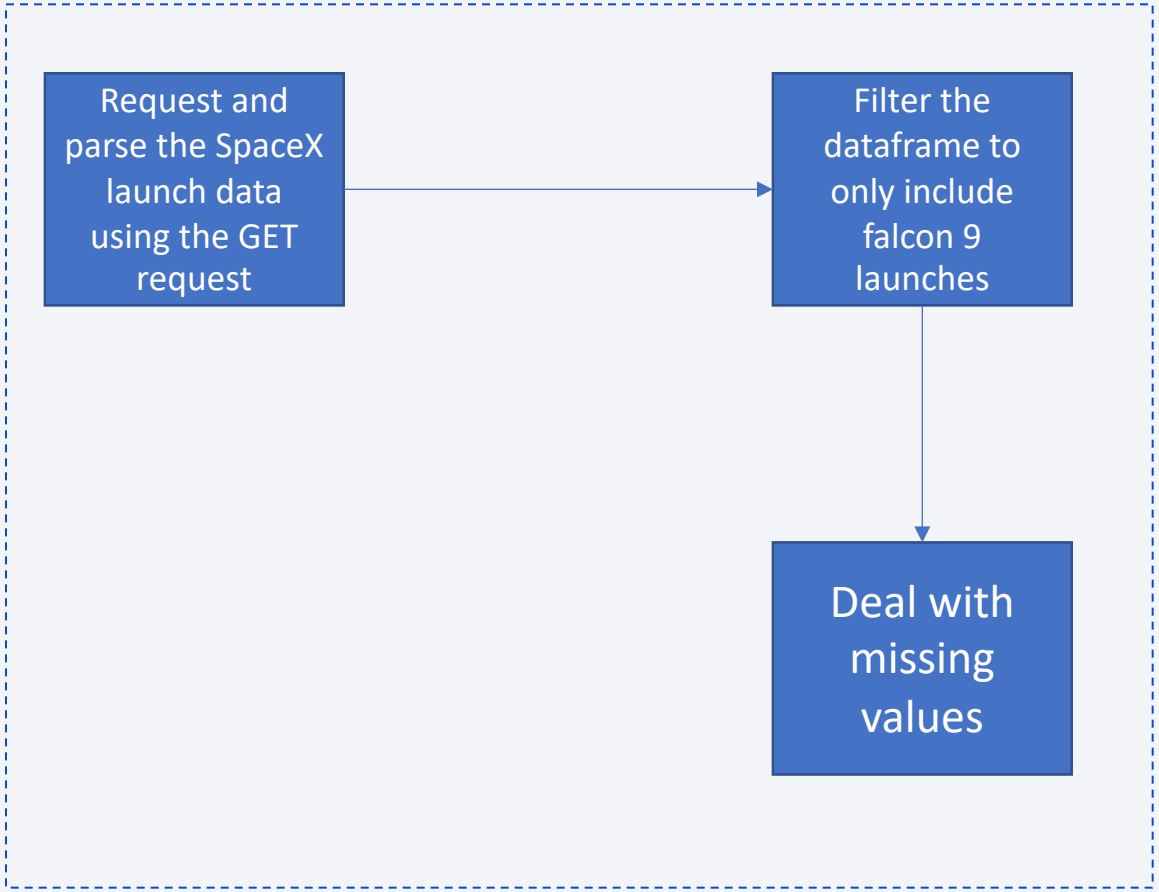
- Data collection methodology:
  - Data was collected using SpaceX API and webscraping.
- Perform data wrangling
  - Creating a landing outcome label from the outcome column.
- Perform exploratory data analysis (EDA) using visualization and SQL
- Perform interactive visual analytics using Folium and Plotly Dash
- Perform predictive analysis using classification models
  - Four different classification models were used. The models were tuned in order to identify the best parameter for each model. The models were evaluated using the accuracy metric.

Data Collection

- Datasets were collected using SpaceX API and from Wikipedia using webscraping techniques.
- Sources:
  - https://api.spacexdata.com/v4/launches/past
  - https://en.wikipedia.org/w/index.php?title=List_of_Falcon_9_and_Falcon_Heavy_launches&oldid=1027686922

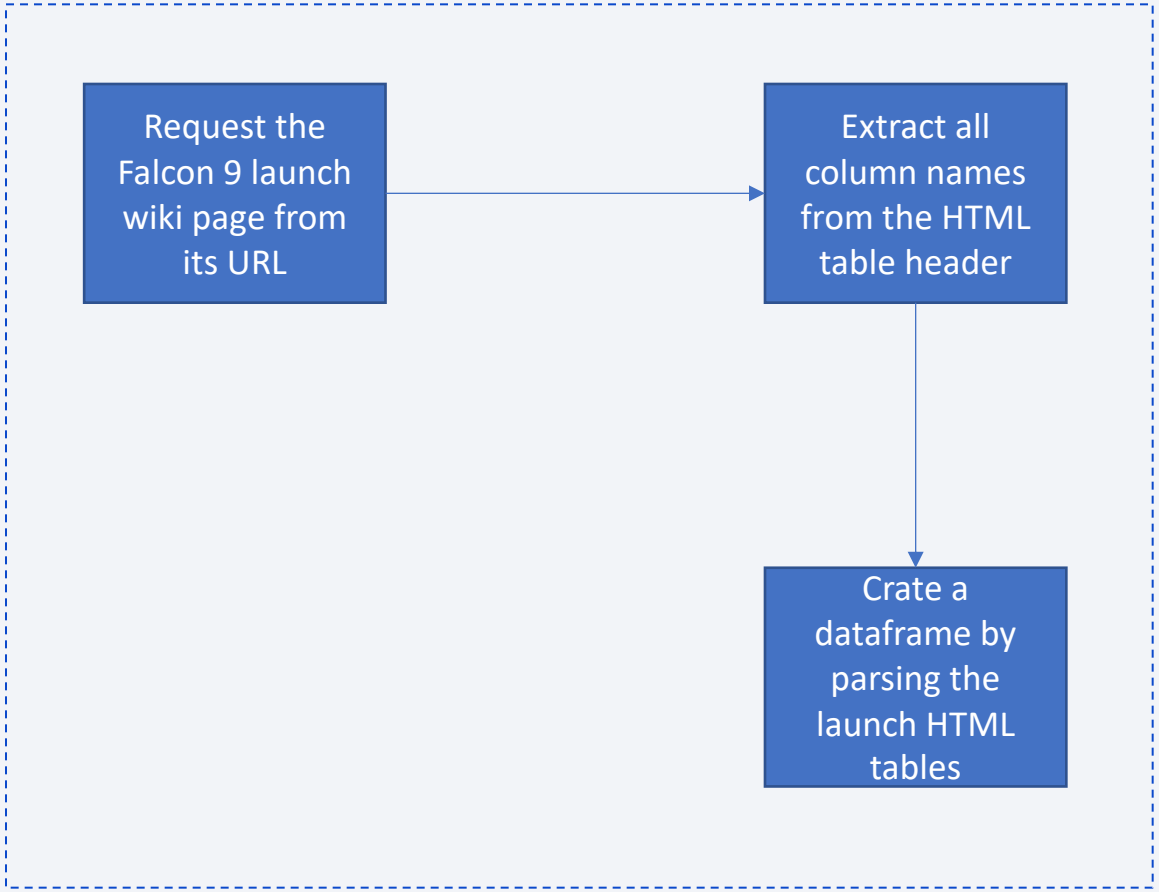
Data Collection – SpaceX API

- Request to the SpaceX API.
- Clean the requested data.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/Data%20Collection%20API.ipynb



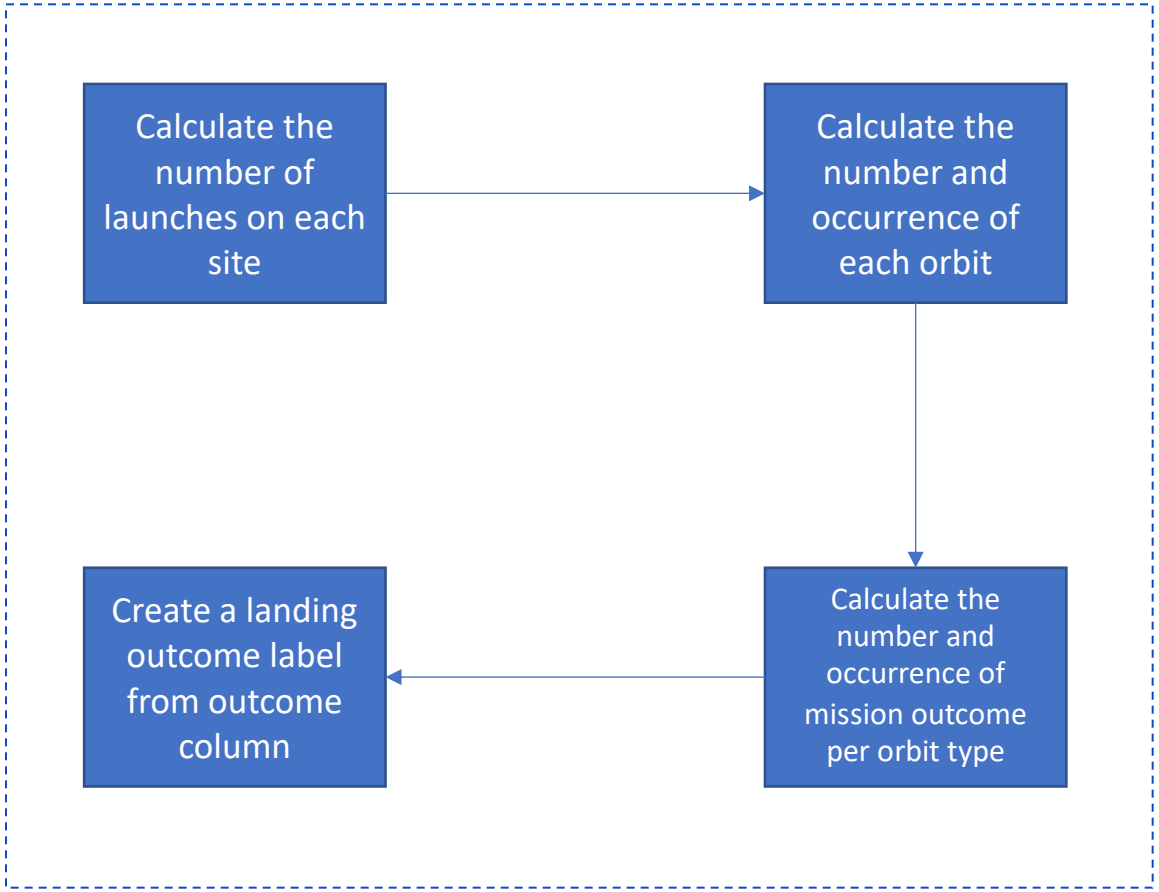
Data Collection - Scraping

- Extract a falcon 9 launch records HTML table from Wikipedia.
- Parse the table and convert it into a pandas dataframe.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/Web scraping.ipynb



Data Wrangling

- Perform exploratory data analysis.
- Determine training labels.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/Data%20Wrangling.ipynb.



EDA with Data Visualization

- Scatter plots and bar plots were used to visualize the relationship between:
  - Flight Number vs. Payload Mass
  - Flight Number vs. Launch Site
  - Payload Mass Vs. Launch Site
  - Flight Number Vs. Orbit
  - Payload Mass Vs. Orbit
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/jupyter-labs-eda-dataviz.ipynb

EDA with SQL

- SQL queries performed:
  - Display the names of the unique launch sites in the space mission.
  - Display 5 records where launch sites begin with the string 'CCA’.
  - Display the total payload mass carried by boosters launched by NASA (CRS).
  - Display average payload mass carried by booster version F9 v1.1.
  - List the date when the first successful landing outcome in ground pad was achieved.
  - List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.
  - List the total number of successful and failure mission outcomes.
  - List the names of the booster versions which have carried the maximum payload mass. Use a subquery.
  - List the failed landing outcomes in drone ship, their booster versions, and launch site names for in year 2015.
  - Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/jupyter-labs-eda-sql-coursera.ipynb

Build an Interactive Map with Folium

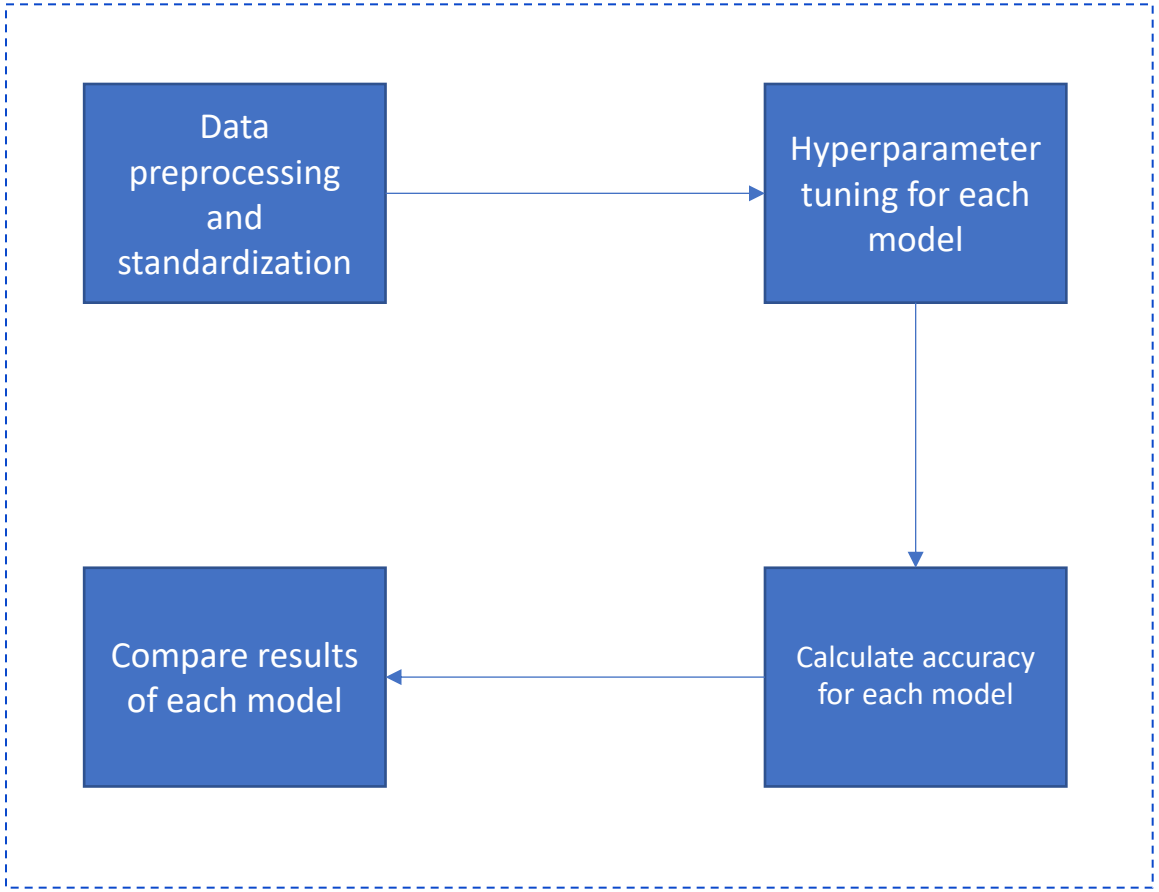
- Map objects used:
  - Markers
    - Markers indicated launch sites on the map.
  - Circles
    - Circles are highlighted areas around specific coordinates.
  - Marker Cluster
    - Marker clusters are used to to simplify a map containing many markers having the same coordinate.
  - Lines
    - Lines are used to indicate distance between two coordinates.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/lab_jupyter_launch_site_location.ipynb

Build a Dashboard with Plotly Dash

- Pie chart used to represent the total success launches by site.
- Scatter plot used to represent the payload range.
- Range slider implemented to allow a user to select a payload mass in a fixed range.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/31ec4ecf2ce5ccdbb520adf27ab51d0adc0150c7/spacex_dash_app.py

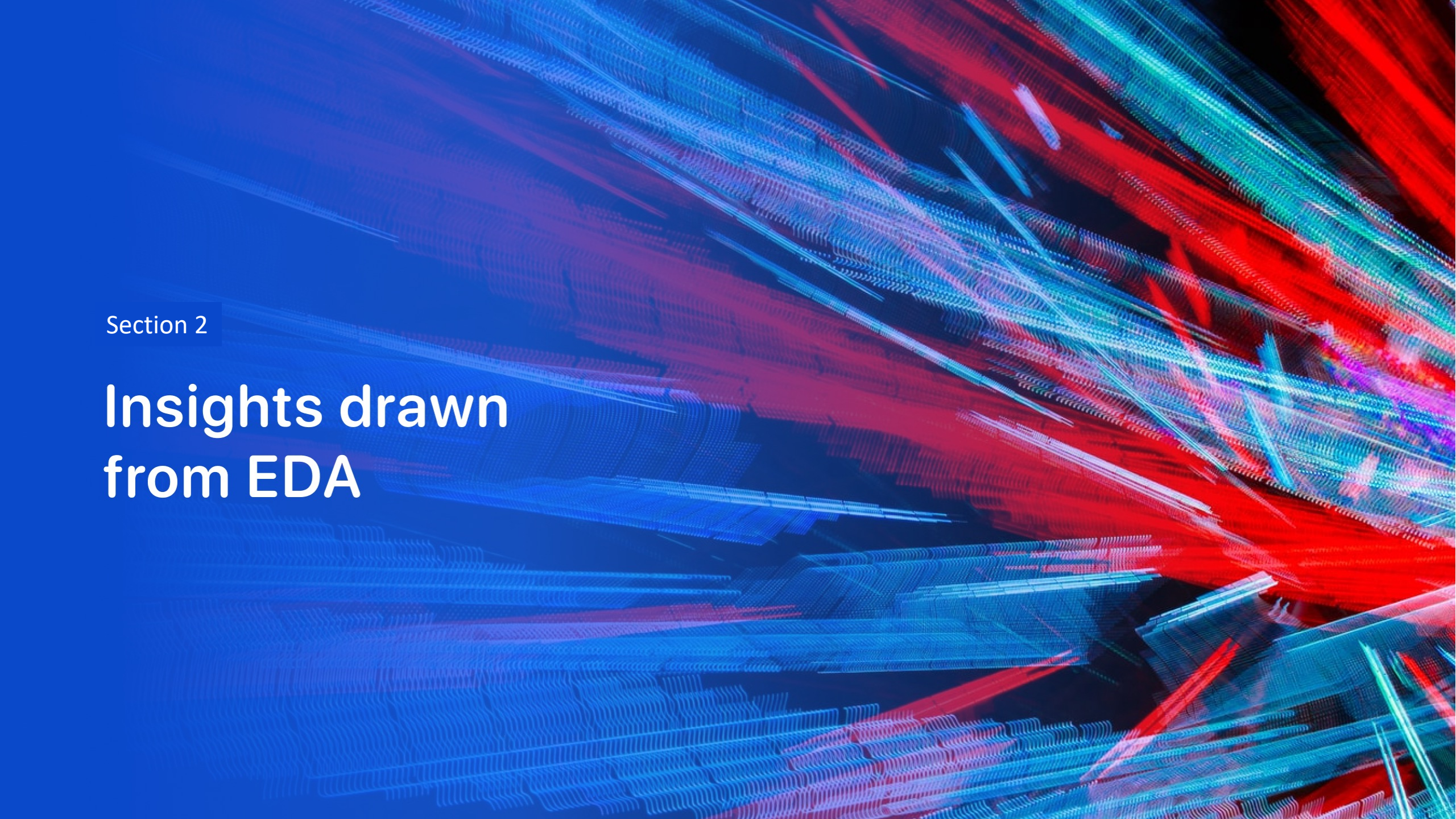
Predictive Analysis (Classification)

- Perform predictive analysis.
- https://github.com/YShanan/Applied-Data-Science-Capstone/blob/767b117584c95c6a46cf06314d3bed564e507672/Machine%20Learning%20Prediction.ipynb



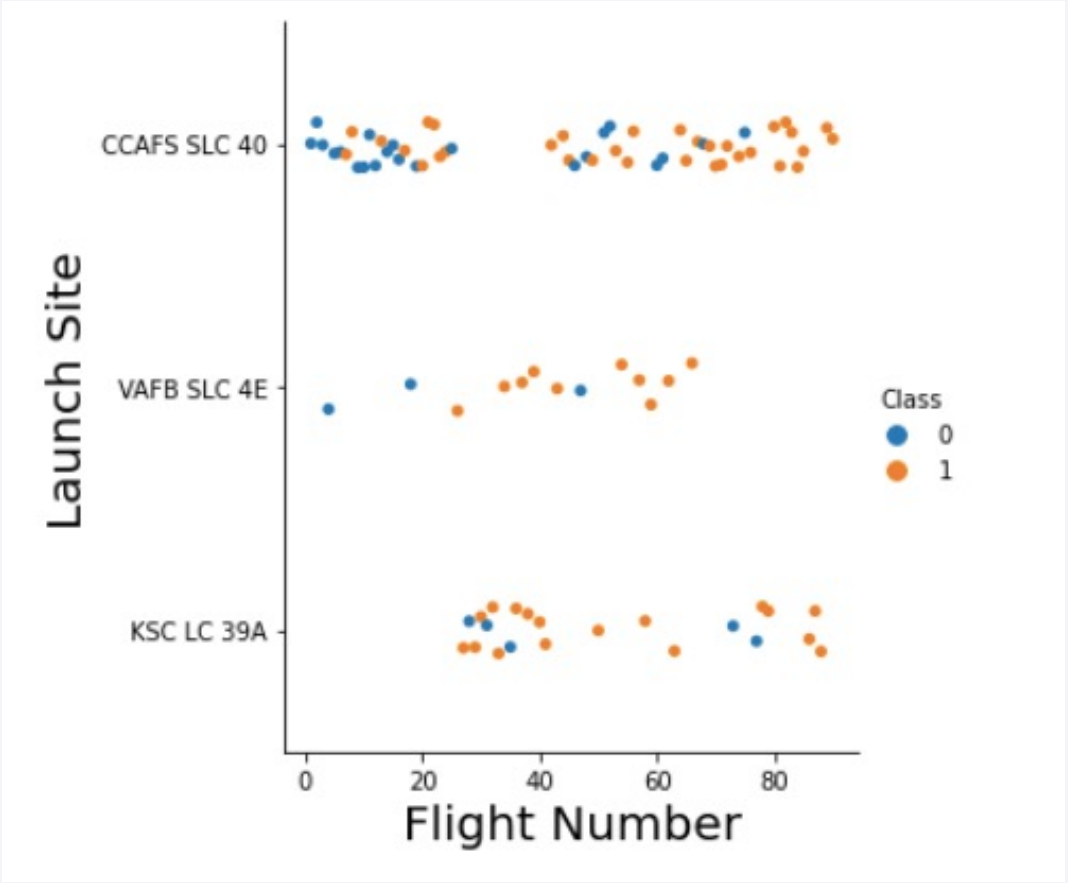
Results

- Exploratory data analysis results
- Interactive analytics demo in screenshots
- Predictive analysis results


Section 2
Insights drawn from EDA

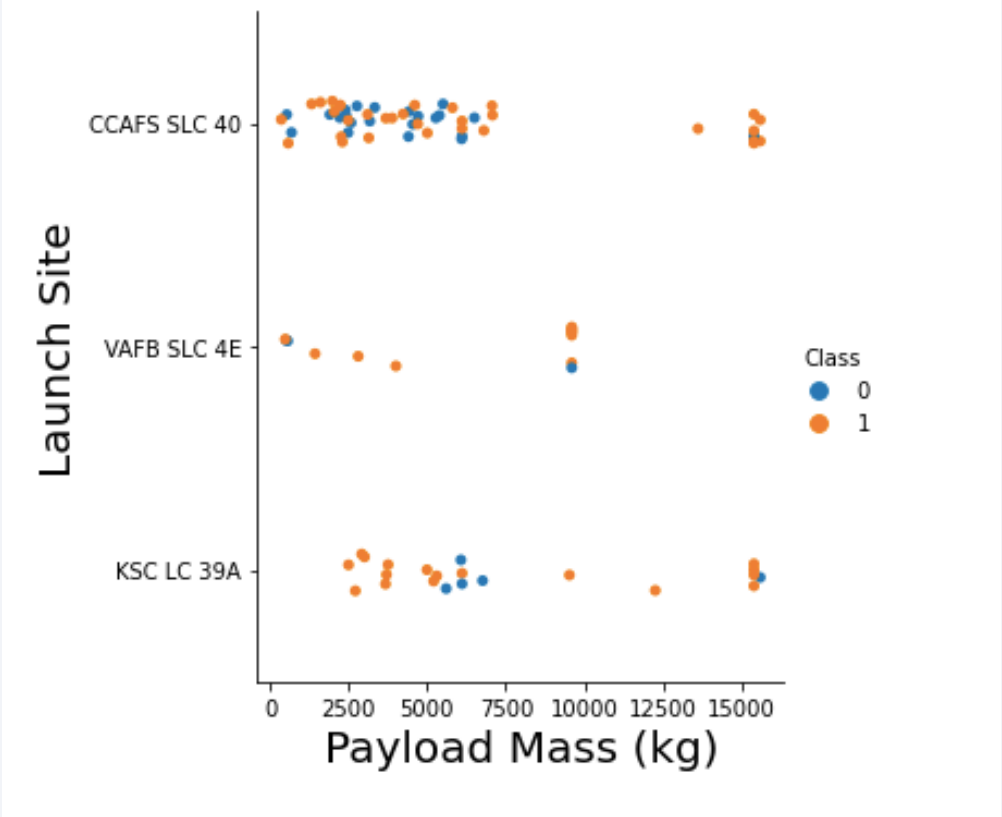
Flight Number vs. Launch Site

- We can notice that as the flight number increases, the success rate increases.



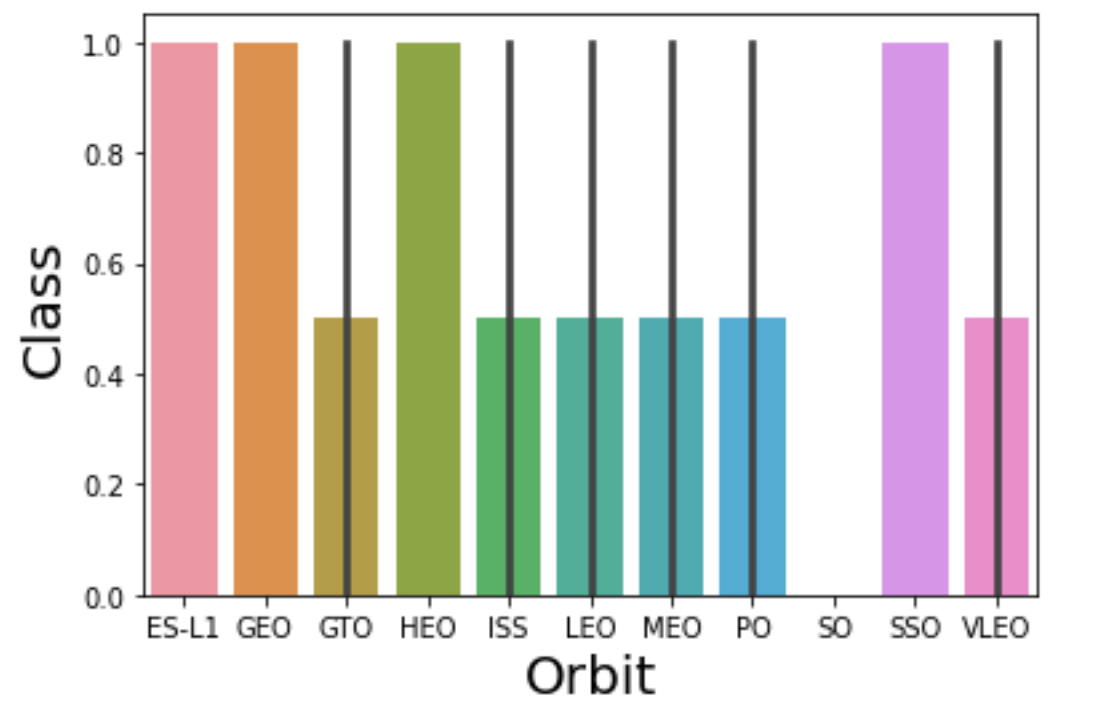
Payload vs. Launch Site

- We can notice that as the payload mass is greater than 9000 kg, the success rate increases.



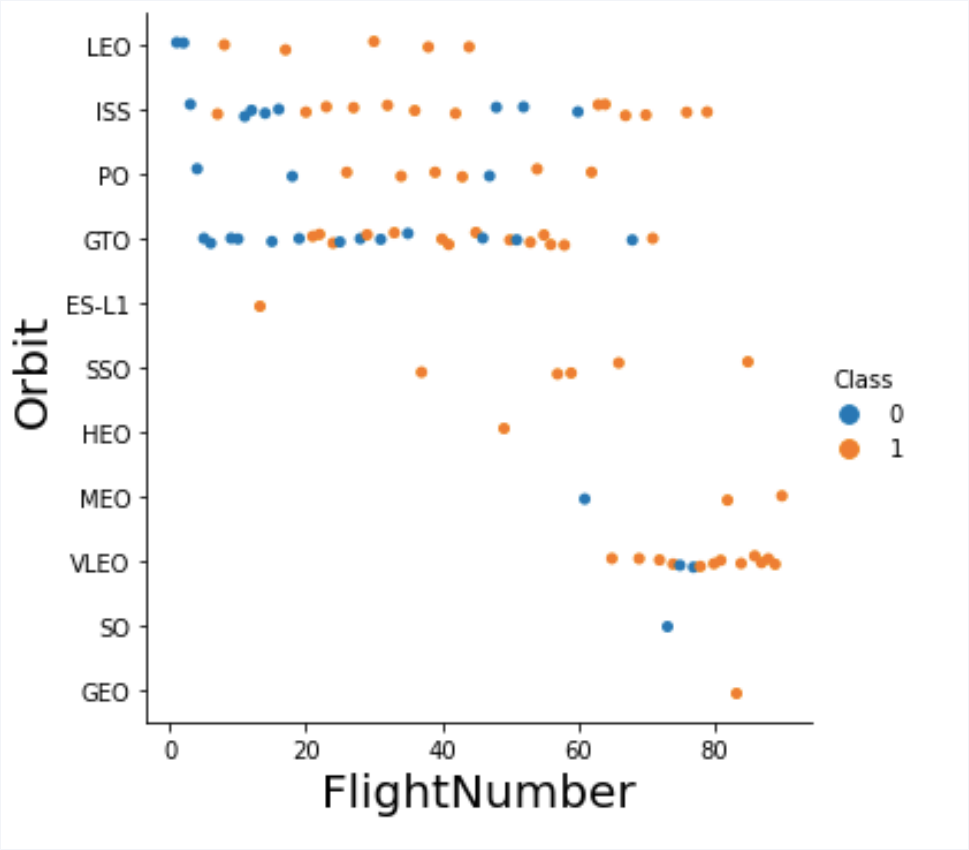
Success Rate vs. Orbit Type

- We can notice that ES-L1, GEO, HEO, and SSO have the highest success rate.



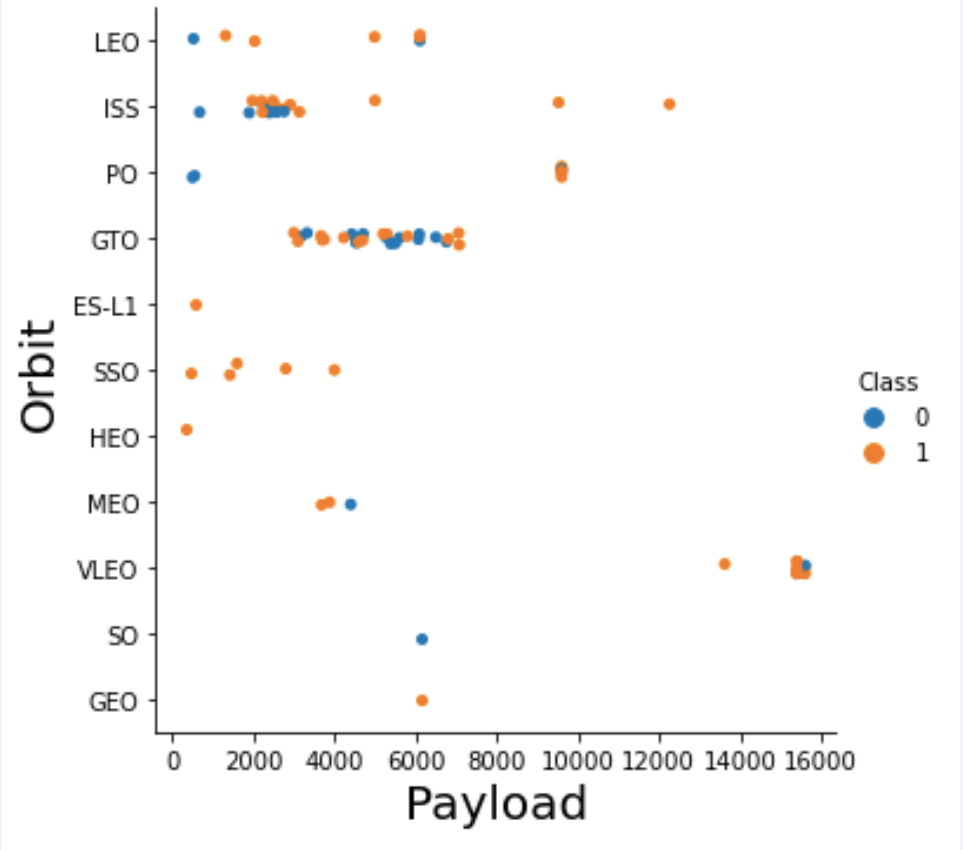
Flight Number vs. Orbit Type

- We can notice for the LEO orbit that as the flight number increases, the success rate increases.



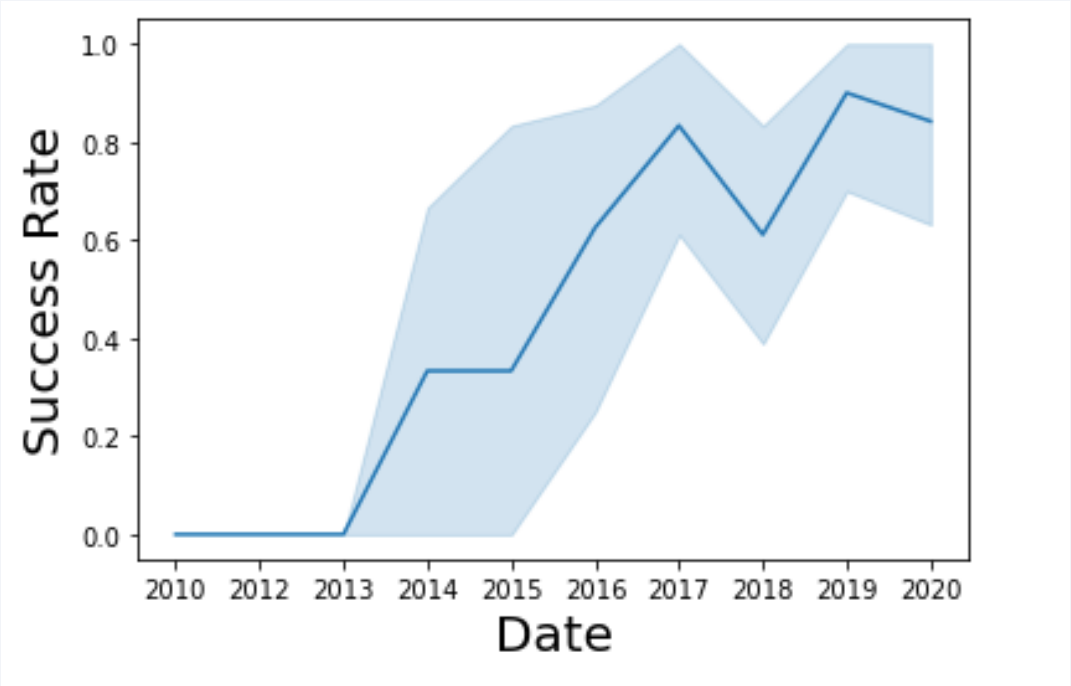
Payload vs. Orbit Type

- We can notice that as the payload mass is greater than 9000 kg, the success rate increases.



Launch Success Yearly Trend

- We can observe that the success rate since 2013 kept increasing till 2020.



All Launch Site Names

- We can observe that there are four different launch sites.

Task 1

Display the names of the unique launch sites in the space mission

In [5]:
```
%sql SELECT DISTINCT LAUNCH_SITE FROM SPACEX ;
```

[redacted]

Done.

Out[5]:

launch_site
CCAFS LC-40
CCAFS SLC-40
KSC LC-39A
VAFB SLC-4E

Launch Site Names Begin with 'CCA'

- The query resulted in only 5 CCAFS LC-40 launch site rows due to the usage of the LIMIT clause.

Task 2

Display 5 records where launch sites begin with the string 'CCA'

```
In [6]: %sql SELECT * FROM SPACEX WHERE LAUNCH_SITE LIKE 'CCA%' LIMIT 5 ;
 * [redacted]
Done.
```

Out[6]:

DATE	time__utc_	booster_version	launch_site	payload	payload_mass__kg_	orbit	customer	mission_outcome	landing__outcome
2010-06-04	18:45:00	F9 v1.0 B0003	CCAFS LC-40	Dragon Spacecraft Qualification Unit	0	LEO	SpaceX	Success	Failure (parachute)
2010-12-08	15:43:00	F9 v1.0 B0004	CCAFS LC-40	Dragon demo flight C1, two CubeSats, barrel of Brouere cheese	0	LEO (ISS)	NASA (COTS) NRO	Success	Failure (parachute)
2012-05-22	07:44:00	F9 v1.0 B0005	CCAFS LC-40	Dragon demo flight C2	525	LEO (ISS)	NASA (COTS)	Success	No attempt
2012-10-08	00:35:00	F9 v1.0 B0006	CCAFS LC-40	SpaceX CRS-1	500	LEO (ISS)	NASA (CRS)	Success	No attempt
2013-03-01	15:10:00	F9 v1.0 B0007	CCAFS LC-40	SpaceX CRS-2	677	LEO (ISS)	NASA (CRS)	Success	No attempt

Total Payload Mass

- The query resulted in outputting the total payload mass carried by boosters launched by NASA (CRS) due to the usage of the SUM clause.

Task 3

Display the total payload mass carried by boosters launched by NASA (CRS)

```
In [7]: %sql SELECT SUM(PAYLOAD_MASS__KG_) FROM SPACEX WHERE CUSTOMER = 'NASA (CRS)'
 * [redacted]
Done.
Out[7]:
1
45596
```

Average Payload Mass by F9 v1.1

- The query resulted in outputting the average payload mass carried by booster version F9 v1.1 due to the usage of the AVG clause.

Task 4

Display average payload mass carried by booster version F9 v1.1

```
In [8]: %sql SELECT AVG(PAYLOAD_MASS__KG_) FROM SPACEX WHERE BOOSTER_VERSION = 'F9 v1.1'
 * [redacted]
Done.
Out[8]:
1
2928
```

First Successful Ground Landing Date

- We can observe the date of the first successful landing outcome in ground pad.

Task 5

List the date when the first successful landing outcome in ground pad was acheived.

Hint:Use min function

```
In [9]: %sql SELECT MIN(DATE) FROM SPACEX WHERE LANDING__OUTCOME = 'Success (ground pad)'
```

[redacted]

```
Done.
```

Out[9]:

1
2015-12-22

Successful Drone Ship Landing with Payload between 4000 and 6000

- We can observe the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000.

Task 6

List the names of the boosters which have success in drone ship and have payload mass greater than 4000 but less than 6000

In [12]:
```
%%sql
SELECT BOOSTER_VERSION FROM SPACEX WHERE LANDING__OUTCOME ='Success (drone ship)' AND
PAYLOAD_MASS__KG_ BETWEEN 4000 AND 6000
```

[redacted]

Done.

Out[12]:

booster_version
F9 FT B1022
F9 FT B1026
F9 FT B1021.2
F9 FT B1031.2

Total Number of Successful and Failure Mission Outcomes

- We can observe the total number of successful and failed mission outcomes due to using the COUNT clause.

Task 7

List the total number of successful and failure mission outcomes

In [13]:
```
%%sql SELECT COUNT(*) FROM SPACEX WHERE MISSION_OUTCOME = 'Success' OR MISSION_OUTCOME = 'Failure (in flight)'
OR MISSION_OUTCOME = 'Success (payload status unclear)'
```

[redacted]

Done.

Out[13]:

1
101

Boosters Carried Maximum Payload

- We can observe the names of the boosters which have carried the maximum payload mass due to using the MAX clause in a sub query.

Task 8

List the names of the booster_versions which have carried the maximum payload mass. Use a subquery

In [19]:
```
%%sql SELECT BOOSTER_VERSION FROM SPACEX WHERE PAYLOAD_MASS__KG_ = (SELECT MAX(PAYLOAD_MASS__KG_) FROM SPACEX)
ORDER BY BOOSTER_VERSION
```

[redacted]

Done.

Out[19]:

booster_version
F9 B5 B1048.4
F9 B5 B1048.5
F9 B5 B1049.4
F9 B5 B1049.5
F9 B5 B1049.7
F9 B5 B1051.3
F9 B5 B1051.4
F9 B5 B1051.6
F9 B5 B1056.4
F9 B5 B1058.3
F9 B5 B1060.2
F9 B5 B1060.3

2015 Launch Records

- We can observe the 2015 failed landing outcomes in drone ship, their booster versions, and launch site names.

Task 9

List the failed landing_outcomes in drone ship, their booster versions, and launch site names for in year 2015

In [22]:
```
%%sql SELECT BOOSTER_VERSION, LAUNCH_SITE, LANDING__OUTCOME FROM SPACEX WHERE
LANDING__OUTCOME = 'Failure (drone ship)' AND DATE LIKE '2015%';
```

[redacted]

```
Done.
```

Out[22]:

booster_version	launch_site	landing__outcome
F9 v1.1 B1012	CCAFS LC-40	Failure (drone ship)
F9 v1.1 B1015	CCAFS LC-40	Failure (drone ship)

Rank Landing Outcomes Between 2010-06-04 and 2017-03-20

- We can observe all the landing outcomes and their count in descending order from the date 2010/6/4 to the date 2017/3/20.

Task 10

Rank the count of landing outcomes (such as Failure (drone ship) or Success (ground pad)) between the date 2010-06-04 and 2017-03-20, in descending order

In [24]:
```
%%sql SELECT LANDING__OUTCOME, COUNT(*) AS Total FROM SPACEX WHERE DATE BETWEEN '2010-06-04' AND '2017-03-20'
GROUP BY LANDING__OUTCOME ORDER BY Total DESC;
```

[redacted]

Done.

Out[24]:

landing__outcome	total
No attempt	10
Failure (drone ship)	5
Success (drone ship)	5
Controlled (ocean)	3
Success (ground pad)	3
Failure (parachute)	2
Uncontrolled (ocean)	2
Precluded (drone ship)	1


Section 3
Launch Sites
Proximities Analysis

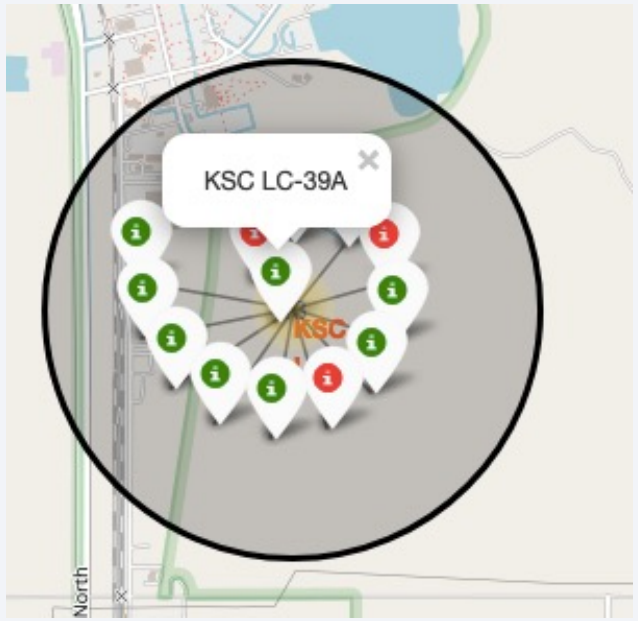
Folium Map – Launch Sites

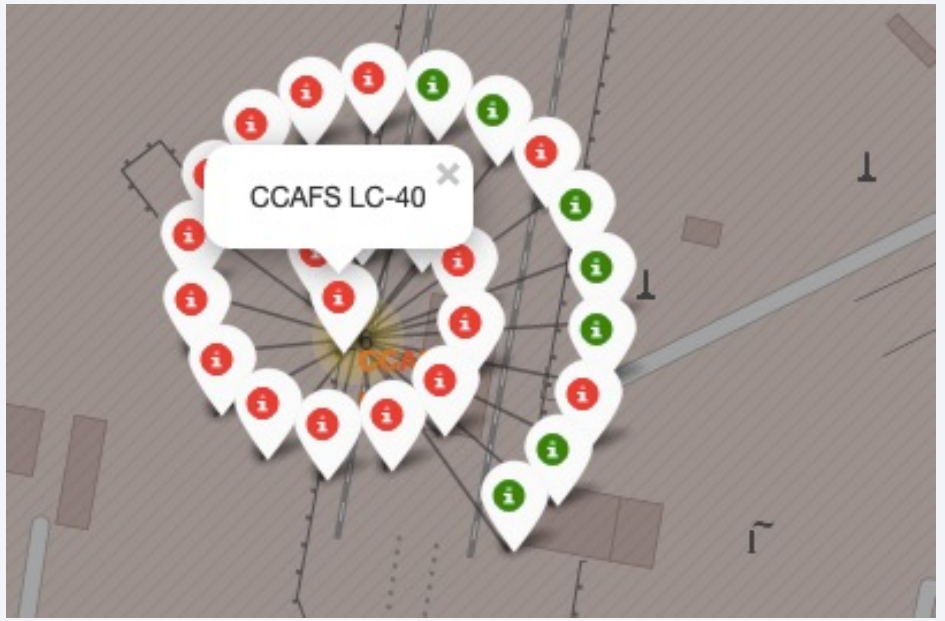
Launch sites marked on the map:

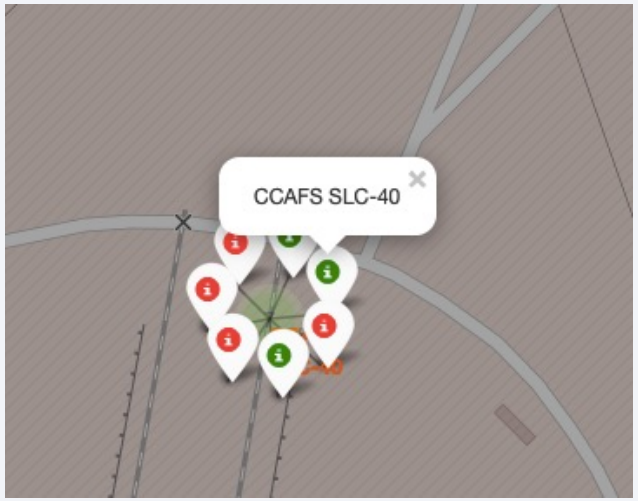
- CCAFS SLC-40
- CCAFS LC-40
- KSC LC-39A
- VAFB SLC-4E

- We can observe that all launch sites are present near the coast.

Folium Map – Color Labeled Launch Outcomes

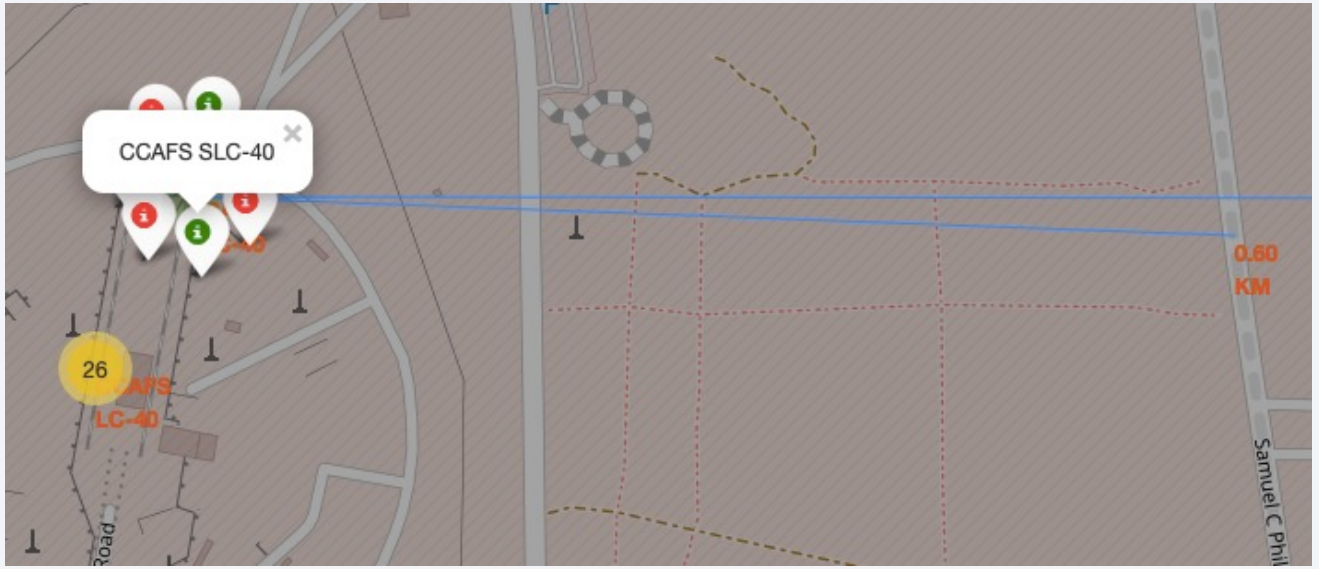




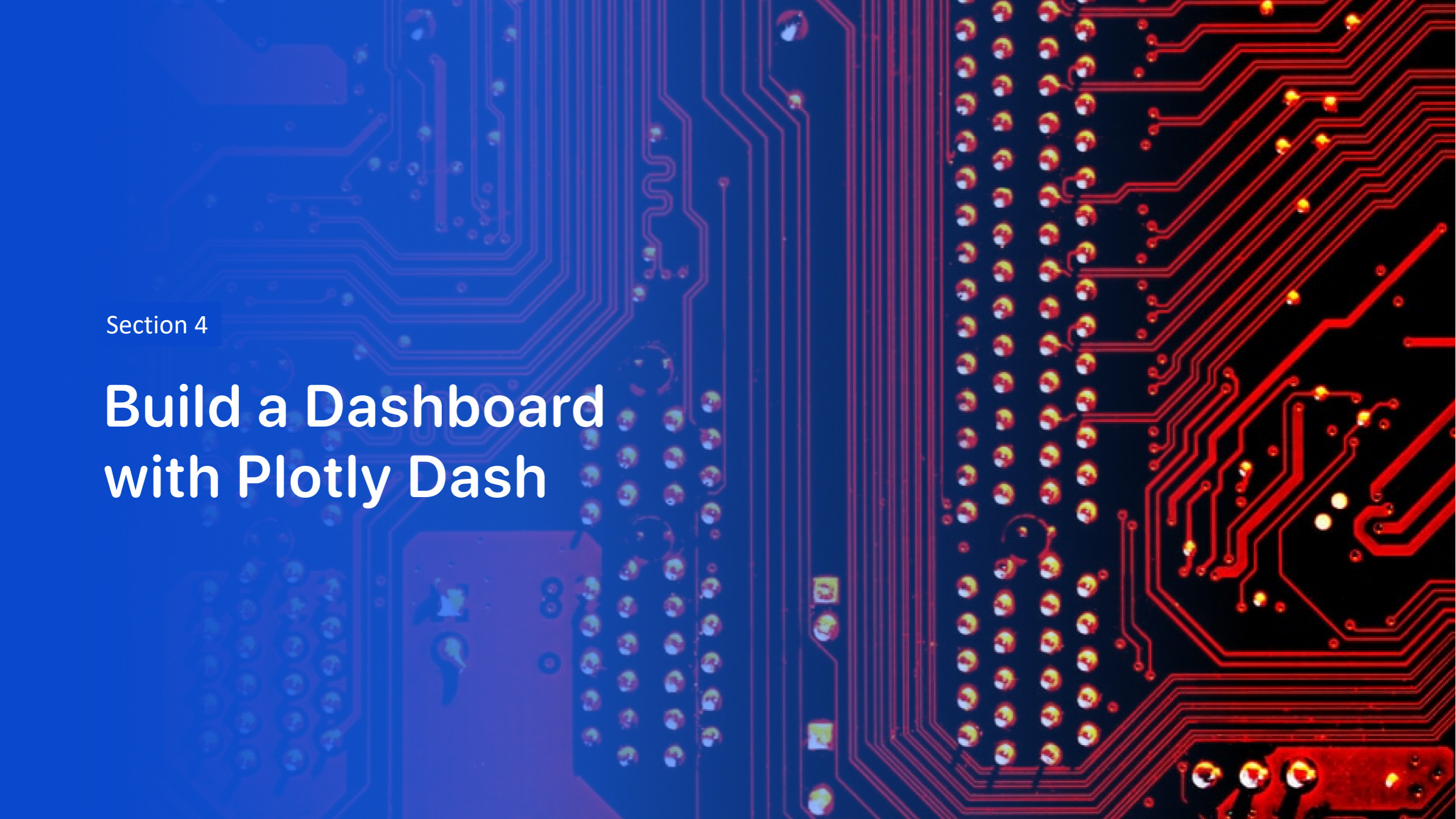


- The green marker indicates a successful launch.
- The red marker indicates a failed launch.
- KSC LC-39A has the highest success rate.

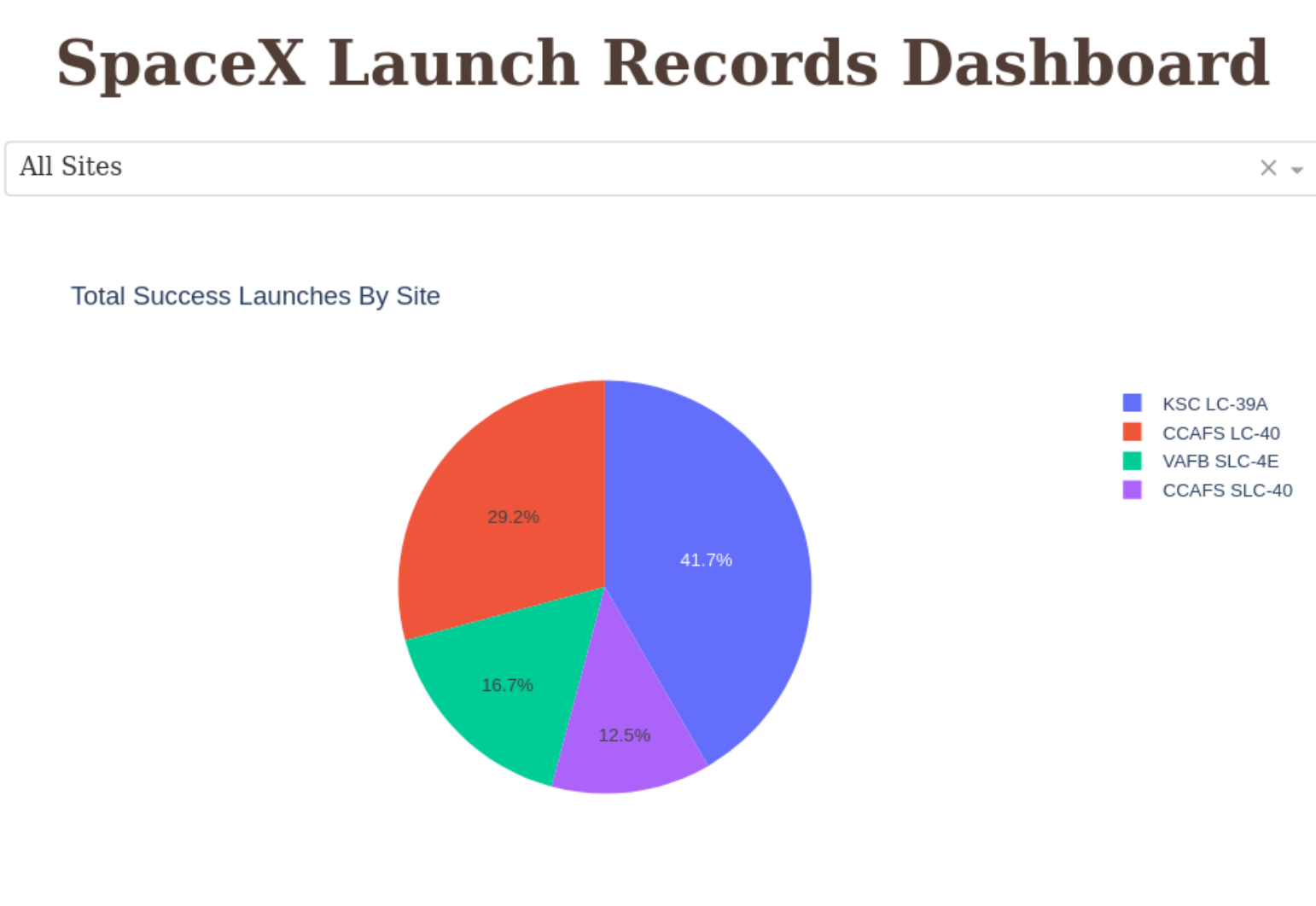
Folium Map – Distance To The Nearest Highway



- The distance between CCAFS SLC-40 and the nearest highway is 0.60 KM.

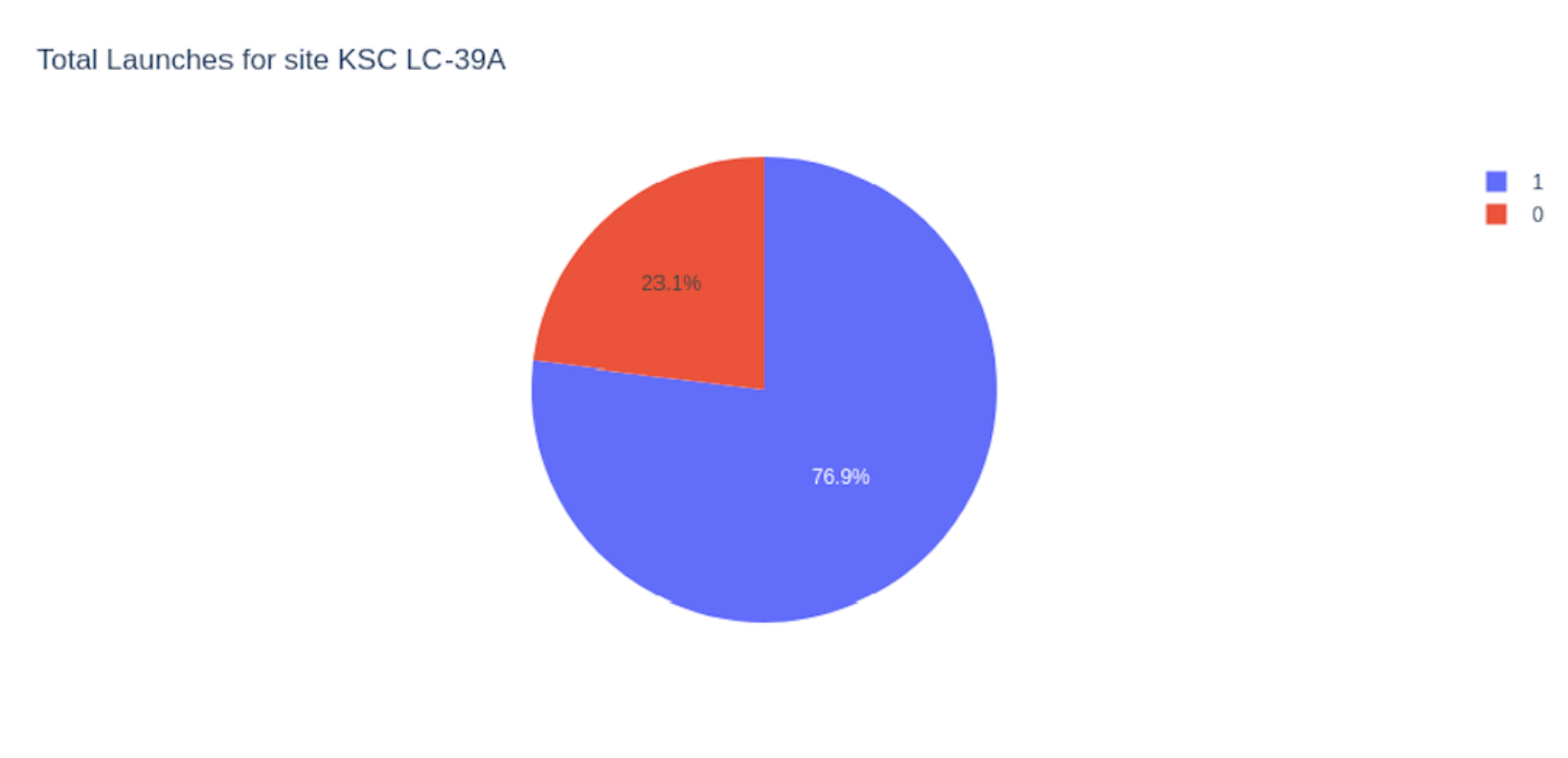

Section 4
Build a Dashboard with Plotly Dash

Dashboard – Total Success Launches By Site



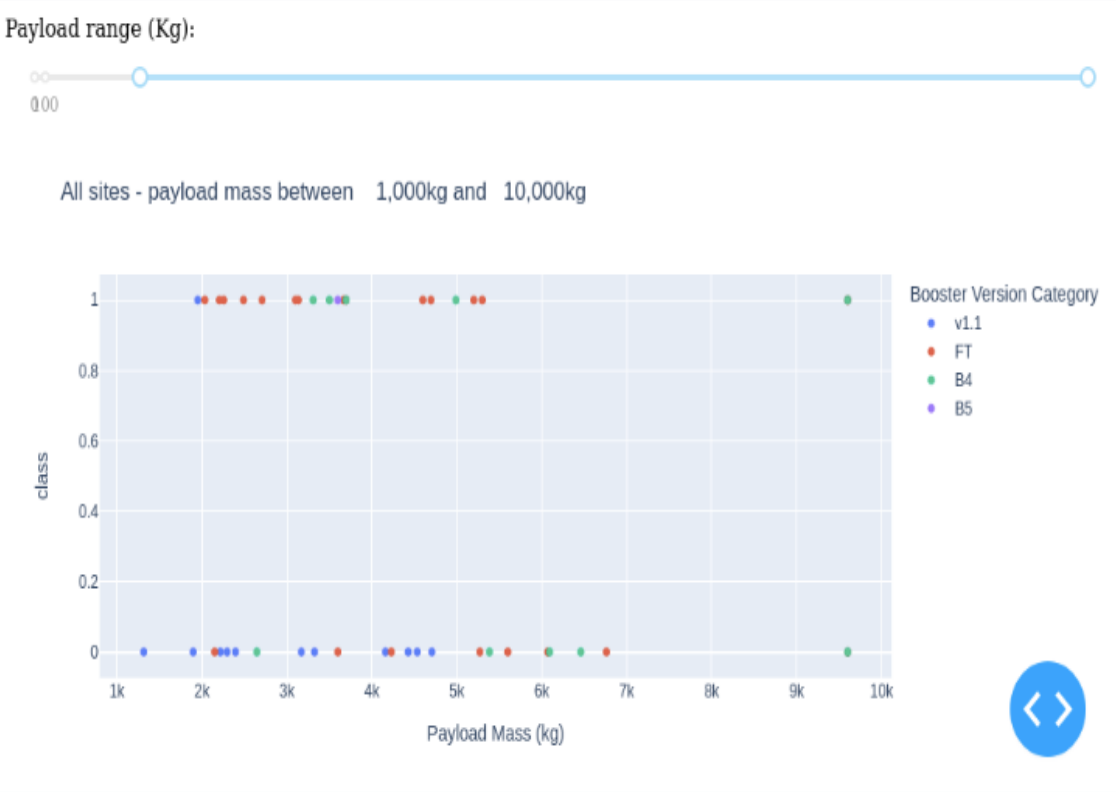
- We can observe that KSC LC-39A has the best success rate.

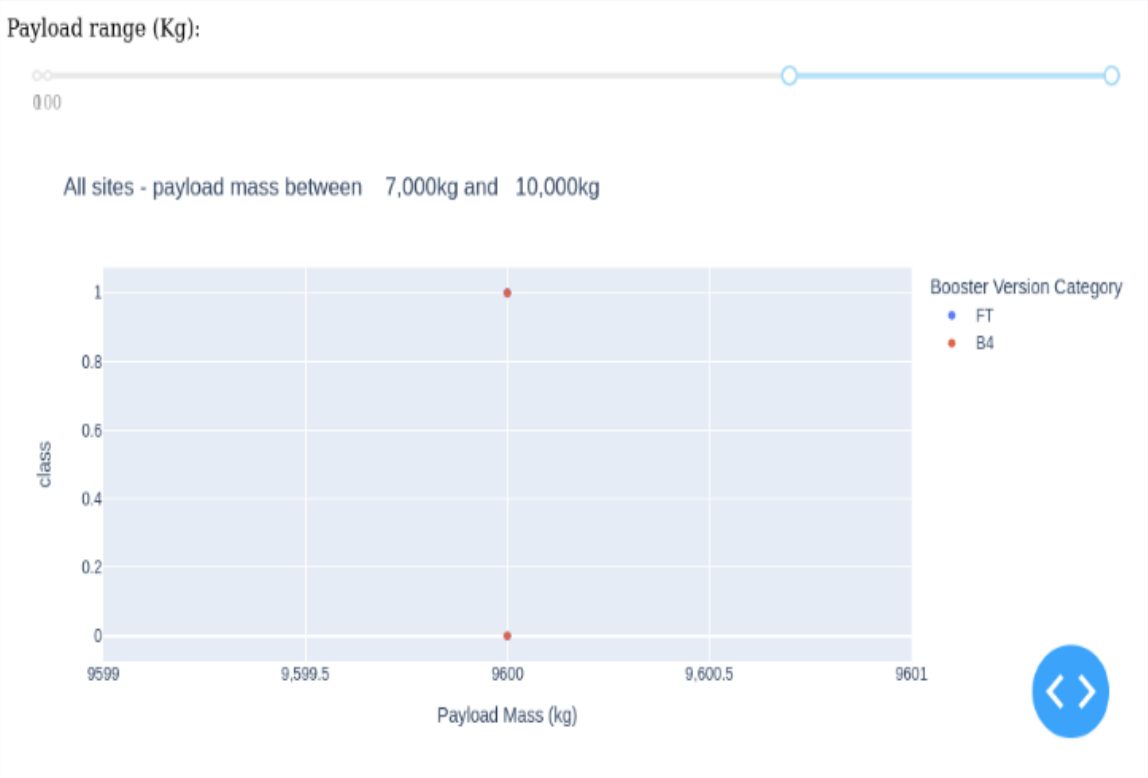
Dashboard – Total Launches for KSC LC-39A



- We can observe that for KSC LC-39A, 76.9% of the total launches are successful.

Dashboard – Payload Vs. Launch Outcome



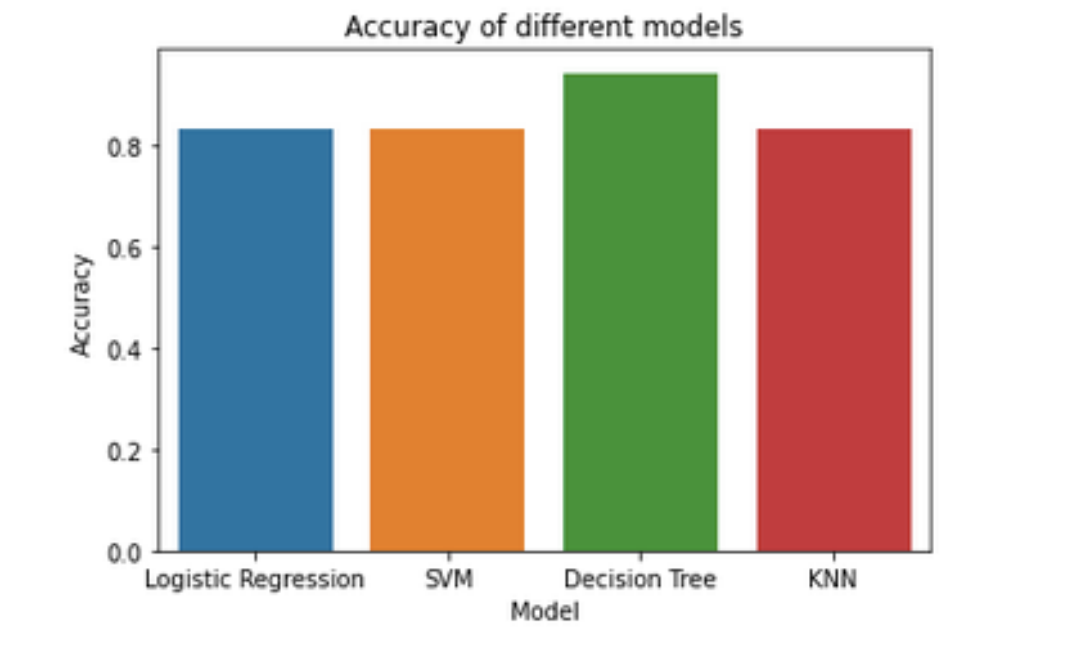


- We can observe that there are a very few number of launches when the payload mass is greater than 7000 KG.

Section 5

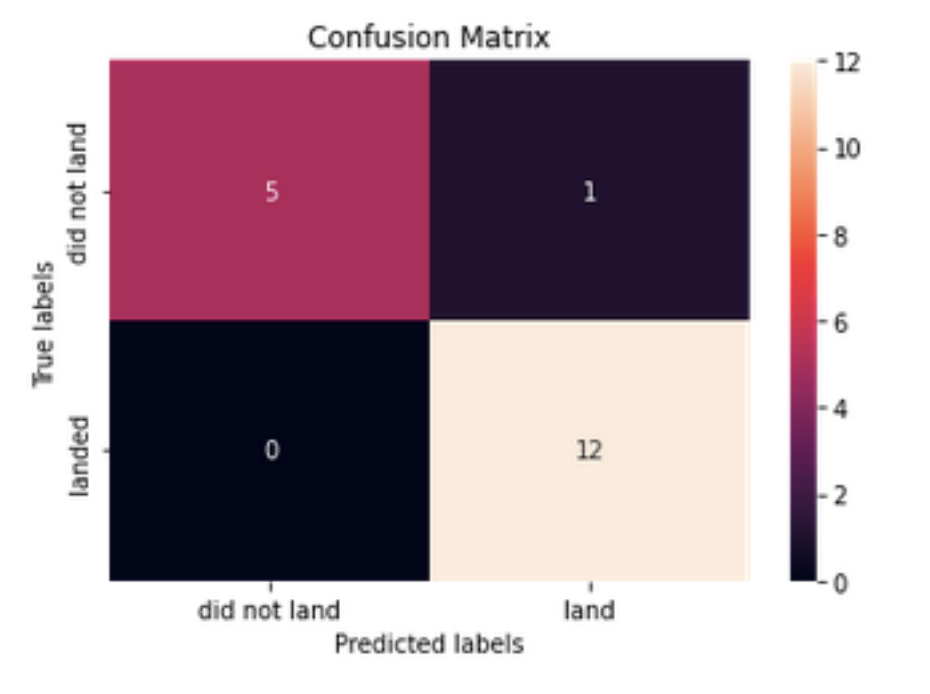
Predictive Analysis (Classification)

Classification Accuracy



- We can observe that the decision tree classifier has the highest accuracy and is the best performing model.

Confusion Matrix



- This is the confusion matrix of the decision tree model.
- We can observe that there are 6 true positives and 5 true negatives.

Conclusions

- The orbits ES-L1, GEO, HEO, and SSO have the highest success rate.
- The success rate since 2013 kept increasing till 2020.
- KSC LC-39A launch site has the best success rate.
- For KSC LC-39A, 76.9% of the total launches are successful.
- The decision tree classifier has the highest accuracy and is the best performing model.

Thank you!